Multivariate Analysen mit zufallsüberlagerten Tabellen aus dem Statistischen Informationssystem des Bundes (STATIS-BUND)

Heer, Georg; Schimpl-Neimanns, Bernhard

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Heer, G., & Schimpl-Neimanns, B. (1992). Multivariate Analysen mit zufallsüberlagerten Tabellen aus dem Statistischen Informationssystem des Bundes (STATIS-BUND). *ZUMA Nachrichten*, *16*(30), 66-94. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-209685

Multivariate Analysen mit zufallsüberlagerten Tabellen aus dem Statistischen Informationssystem des Bundes (STATIS-BUND)

Von Georg Heer und Bernhard Schimpl-Neimanns

Das Statistische Informationssystem des Bundes (STATIS-BUND) bietet Nutzern außerhalb der amtlichen Statistik unter bestimmten Voraussetzungen die Möglichkeit, per Online-Anschluß amtliche Mikrodaten nach eigenen Wünschen auszuwerten. Die Nutzer erhalten Fallzahltabellen, die aus Geheimhaltungsgründen mit Zufallsvariablen überlagert sind. Ein Vergleich der Analysen von überlagerten Tabellen mit Analysen der Originaltabellen am Beispiel von Mikrozensus-Daten zeigt keine wesentlichen Verzerrungen in den Ergebnissen. Lediglich sehr schwach besetzte Tabellenfelder verursachen Unterschiede in Teilergebnissen. Des weiteren werden Möglichkeiten diskutiert, den Überlagerungsfehler bei multivariaten Analysen zu berücksichtigen. Die Untersuchungen über die Auswirkungen dieser Überlagerung auf die Ergebnisse multivariater Analysen wurden in Zusammenarbeit zwischen dem Statistischen Bundesamt und ZUMA durchgeführt. Georg Heer ist Referent in der Gruppe *Statistisches Informationssystem des Bundes* im Statistischen Bundesamt.

1. Einleitung

Mit STATIS-BUND bietet das Statistische Bundesamt auch Nutzern außerhalb der amtlichen Statistik Zugang zu Ergebnissen aus der Bundesstatistik. Das EDV-gestützte Informationssystem enthält eine Vielzahl von Zeitreihen und Strukturdaten in Tabellenform sowie komplexe Auswertungs- und Analysemöglichkeiten.

Daneben besteht unter gewissen Voraussetzungen die Möglichkeit, Auswertungen aus Mikrodaten nach eigenen Vorgaben zu erhalten, die der gesetzlich vorgeschriebenen statistischen Geheimhaltungspflicht genügen. Dies erfolgt, indem im Online-Verfahren anonymisierte Fallzahltabellen aus dem Einzelmaterial erstellt werden. Dabei spezifizieren die Nutzer die Tabellen nach eigenen Anforderungen und sind somit nicht an die amtlichen Klassifikationen gebunden. Bei dieser Form des indirekten Zugriffs auf amtliche Mikrodaten wird kein Einzelmaterial weitergegeben. Der vorliegende Aufsatz betrifft damit nicht die Weitergabe eines sogenannten "faktisch anonymisierten Mikrodatenfiles" an die Wissenschaft nach Paragraph 16 Abs. 6 BStatG (vgl. hierzu den Aufsatz von Heike Wirth in diesem Heft).

Auch bei der Weitergabe von Tabellen dürfen keine Einzelangaben offenbar werden. Hier wurden bisher hauptsächlich manuelle Anonymisierungsverfahren angewendet (Streichung von Zellenbesetzungen kleiner als drei,

Aggregation, Klassenbildung etc.), die jedoch aufwendig, fehleranfällig und mit einem Informationsverlust verbunden sind. Diese Nachteile sollen in STATIS-BUND durch ein Verfahren der automatischen Überlagerung von Tabellen mit ganzen Zufallszahlen so weit als möglich vermieden werden.

Bei der Abschätzung des durch die Überlagerung entstehenden Fehlers wurde bereits festgestellt, daß dieser bei der freien Hochrechnung einzelner Tabellenfelder nicht wesentlich über dem beim Mikrozensus üblichen Stichprobenfehler liegt (Kühn/Pfrommer/Schrey 1984). Auf multivariate Analysen ist dieser Befund jedoch nicht ohne weiteres übertragbar, da die Struktur der Tabelle in die bisherigen Fehlerrechnungen nicht einging. Um die Auswirkungen der Überlagerung in STATIS-BUND auf multivariate Analysen zu untersuchen, vergleichen wir in diesem Aufsatz die Ergebnisse von Logit-Modellen für Originaltabellen mit den Ergebnissen für überlagerte Tabellen. Um eine möglichst praxisrelevante Auswertungssituation zu haben, wurden Fragestellungen der schichtspezifischen Bildungsungleichheit untersucht. Methodisch konzentrieren wir uns dabei auf eine spärlich besetzte Tabelle, bei der zu vermuten ist, daß sich die Überlagerung besonders kritisch auswirkt. Darüber hinaus geben wir Hinweise, wie die Auswirkungen des Überlagerungsfehlers möglichst klein gehalten werden können.[1)]

2. Das Statistische Informationssystem des Bundes

2.1 Leistungsumfang

STATIS-BUND enthält statistische Ergebnisse in Form von Zeitreihen und Tabellen mit Strukturdaten aus allen Bereichen der amtlichen Statistik. Das Spektrum des Angebots in den circa 900000 Zeitreihen und in den Strukturtabellen reicht von Daten zu Bevölkerung, Erwerbstätigkeit und Wahlen über Betriebs- und Unternehmensstatistiken bis hin zu Außenhandelszahlen und Volkswirtschaftlichen Gesamtrechnungen. Einen Überblick über das aktuelle Angebot gibt das jährlich vom Statistischen Bundesamt herausgegebene *Datenbestandsverzeichnis*. Die Statistiken sind innerhalb des Systems fachlich und technisch ausführlich beschrieben. Die Dokumentation enthält alle Informationen, die bei der Auswahl und Verwendung der Daten sowie bei der Beurteilung ihrer Qualität erforderlich sind.

Neben dem Angebot an allgemein zugänglichen aggregierten Daten besteht unter bestimmten Voraussetzungen die Möglichkeit, amtliche Mikrodaten für die Weiterverarbeitung in einer anonymisierten Form zu nutzen. Diese Möglichkeit ist unter Punkt 2.2 näher beschrieben.

Ein Nutzer kann Online auf Daten, Analyse- und Auswertungsverfahren in STATIS-BUND zugreifen, sofern er über einen Datex-P-Hauptanschluß verfügt und einen entsprechenden Vertrag mit dem Statistischen Bundesamt abgeschlossen hat. Diese Nutzungsart ist vor allem dann sinnvoll, wenn die vielfältigen Möglichkeiten des Systems intensiv genutzt werden. Eine komfortable Benutzersprache erlaubt dem Anwender die Weiterverarbeitung von Systemdaten sowie eigener Daten mit den folgenden Mitteln:

Auswertungssystem: Verfahren zur Tabellenerstellung einschließlich hierarchischer Auswertung, Mischen, Sortieren und Bearbeiten von Materialien mittels Rechenoperationen sowie Druckaufbereitung.

Analysesystem: Mathematisch-Statistische Analysemethoden, z.B. Lösen von Gleichungssystemen, Regression, Interpolation, Prognose, Zeitreihenanalyse, Bevölkerungsanalyse und -prognose, Varianzanalyse, Faktorenanalyse, Lag-Untersuchungen.

Grafiken: grafische Darstellung von Ergebnissen, Kurven-, Balken- Kreis- oder Tortendiagramme, Deutschlandkarte, Europakarte, Ausgabe in vordefinierter oder in beliebiger Form.

Neben dem Online-Anschluß gibt es noch eine Reihe weiterer Möglichkeiten, Daten aus STATIS-BUND zu erhalten. Für den Bezug kleinerer Datenmengen oder die gezielte Auswahl spezieller Zeitreihen bietet sich der *Diskettenservice* an. Die Disketten können einmalig oder im Rahmen eines Jahresabonnements monatlich, viertel- bzw. halbjährlich bezogen werden. Umfangreichere Datenbestände können per *Magnetband* geliefert werden.[2)]

2.2 Erstellung zufallsüberlagerter Fallzahltabellen aus Einzeldaten

Wie eingangs erwähnt, besteht darüber hinaus die Möglichkeit, Auswertungen aus Mikrodaten zu erhalten. Der Benutzer beschreibt die gewünschte Tabelle in der Benutzersprache des Systems und erteilt durch ein Kommando den Auftrag zur Tabellenerstellung. Aufgrund dieses Auftrags wird die Berechtigung des Benutzers automatisch geprüft und das Originalmaterial zu einer Tabelle ausgezählt, auf die der Benutzer jedoch keinen Zugriff hat. Diese *Originaltabelle* verbleibt im Statistischen Bundesamt für Kontrollzwecke. Aus ihr wird in einem zweiten Schritt eine weitere Tabelle erzeugt, indem zu jeder Fallzahl der Originaltabelle eine ganzzahlige

Zufallszahl addiert wird. Die verwendeten Zufallszahlen sind über mehrere verschiedene Tabellen hinweg annähernd normalverteilt mit Mittelwert Null und Varianz Drei.[3] Die mit Zufallszahlen *überlagerte Tabelle* wird dem Benutzer zur Verfügung gestellt, der damit eine Tabelle nach seinen Vorgaben erhält, die sich von der Originaltabelle in jedem Feld zufallsabhängig geringfügig unterscheidet. Hat die Tabelle viele schwach besetzte Felder, so wirkt sich die Überlagerung naturgemäß relativ stark aus. Der Benutzer kann dann durch gezielte Änderungen der Spezifikationen, zum Beispiel durch das Zusammenfassen von Gliederungspositionen, und einen erneuten Auftrag versuchen, eine aussagefähigere Tabelle zu erhalten.

Das Verfahren hat den Vorteil, daß es für den Anwender keinerlei Einschränkungen hinsichtlich Anzahl und Spezifikation der zu erstellenden Tabellen gibt. Dafür müssen Genauigkeitsverluste in den Einzelfeldern hingenommen werden. Die Überlagerung ist so konstruiert, daß durch Differenzbildung die wahre Größe eines geheimzuhaltenden Wertes nicht ermittelt werden kann. Für eine konkrete Tabelle sind demzufolge die Zufallszahlen nicht normalverteilt mit Mittelwert Null und Varianz Drei. Diese Aussage gilt nur näherungsweise für viele verschieden aufgebaute Tabellen.

Die Überlagerung hat zur Folge, daß in einer Tabelle die Summe von überlagerten Werten in der Regel nicht mit der überlagerten Summe übereinstimmt. Zur Bestimmung einer möglichst exakten Summe sollte diese mit ausgezählt werden und nicht nachträglich aus überlagerten Werten errechnet werden, da sonst die Varianz des Zufallsfehlers steigt (vgl. dazu Abschnitt 3.1).

Die Güte der überlagerten Einzelwerte wurde bereits durch eine Fehlerbetrachtung anhand von Auswertungen des Mikrozensus untersucht (Kühn/Pfrommer/Schrey 1984). Dazu wurde der einfache absolute und der relative Standardfehler für verschiedene Fallzahlen in einer Mikrozensustabelle mit und ohne Überlagerung geschätzt.[4] Für kleine Fallzahlen ist der Unterschied im relativen Standardfehler noch relativ groß (210 Prozent mit Überlagerung gegenüber 160 Prozent ohne Überlagerung bei einer Besetzungszahl von eins, hochgerechnet 100), er sinkt jedoch schnell. Bereits bei einer Besetzungszahl von zehn beträgt der relative Standardfehler 54 Prozent gegenüber 51 Prozent in der Originaltabelle. Für eine Fallzahl von 50 liegen die Werte mit 22,8 Prozent bzw. 22,6 Prozent praktisch zusammen. Bei der Überlagerung entsteht also nur bei denjenigen Fallzahlen ein bedeutender zusätzlicher Fehler, die schon wegen ihres großen Stichprobenfehlers nicht gesichert sind. Diese Überlegungen gelten jedoch nur für ein einzelnes Tabellenfeld. Wie sich die Überlagerung mit Zufallszahlen auf die

Gesamtstruktur einer Tabelle auswirkt, wird in den Abschnitten drei und vier untersucht.

3. Einführung in das Problem

In diesem eher didaktisch orientierten Abschnitt werden die Verzerrungen einfacher statistischer Kennzahlen von Tabellen, deren Zellenbesetzungen mit Zufallsvariablen überlagert sind, aufgezeigt. Die dabei verwendeten synthetischen Daten sind in der Tabelle 3.1 beschrieben.

Tabelle 3.1: Anzahl der Schüler nach Schulart und sozialer Herkunft (fiktive Werte)

Schulart	Mittel-schicht	Unter-schicht	Zusammen
weiterführend	30	25	55
Hauptschule	20	40	60
Zusammen	50	65	115

Zu jedem Tabellenfeld dieser Ausgangstabelle wird, abweichend von der Überlagerung in STATIS-BUND, eine normalverteilte Zufallszahl mit Mittelwert Null und Varianz Eins addiert. Die Zufallszahlen sind stochastisch unabhängig.

3.1 Randsummen und Anteilswerte

Für den Mittelwert und die Varianz einer Randsumme, die aus den Besetzungszahlen von K Zellen ($X_1, \ldots, X_K$) gebildet werden, gilt:

(3.1) $E(\Sigma X_k) = \Sigma E(X_k)$

(3.2) $Var(\Sigma X_k) = \Sigma var(X_k) = K\sigma^2$

da die Zufallsfehler voneinander unabhängig sind. Zählt man jedoch die Randverteilungen zusätzlich zu den Zellen aus, weisen diese nur jeweils eine Varianz von σ^2 auf, sind also weniger fehlerbehaftet als die durch Summierung der Zellen gebildeten Randsummen.

Um die Verzerrung eines Anteilswertes festzustellen, gehen wir von der tatsächlichen Zellenbesetzung $\mu(x)$ und der tatsächlichen Randsumme $\mu(n)$ aus. Die überlagerte Zellenbesetzung X besitzt den Mittelwert $\mu(x)$ und die Varianz $\sigma^2(n)$. Die Randsumme N mit Mittelwert $\mu(n)$ und Varianz $\sigma^2(n)$ kann entweder die einfach überlagerte Randsumme sein oder sich als Summe überlagerter Einzelzellen ergeben. Für den *erwarteten Anteilswert*, der besonders bei kleinen Zellenbesetzungen überschätzt wird, erhalten wir näherungsweise (vgl. Punkt 1 im Anhang):

$$(3.3) \quad E\left(\frac{X}{N}\right) \approx \frac{\mu_x}{\mu_n}\left(1 + \frac{\sigma_n^2}{\mu_n^2} - \frac{\sigma_{xn}^2}{\mu_x \mu_n}\right)$$

Da in der Näherungsformel nur die ersten beiden Momente berücksichtigt werden, wurden Monte-Carlo Simulationen zum Stichprobenumfang 1000 durchgeführt. Eine Verzerrung durch die Überlagerung konnte für die Beispieltabelle praktisch nicht festgestellt werden: Der Anteil an Schülern auf weiterführenden Schulen lag im Mittel bei 0,6 bzw. 0,3845 bei einer empirischen Varianz von 0,0141 bzw. 0,0113.

Die Varianz eines Anteilswertes $p=\mu_x/\mu_n$ oder eines Tabellenfeldes μ_x beträgt nach dem Binomialansatz

$$\sigma^2(p)=p(1-p)/\mu_n = \mu_x/\mu_n^2(1-\mu_x/\mu) \text{ bzw. } \sigma^2(\mu_x) = p(1-p)\,\mu_n = \mu_x - \mu_n^2/\mu_n .$$

Beide Maße sind infolge des Anonymisierungsverfahrens verzerrt. Die *Erwartungswerte der geschätzten Varianzen* lassen sich auch hier näherungsweise ermitteln:

$$(3.4) \quad E(\hat{\sigma}_p^2) \approx \frac{\mu_x}{\mu_n^2}\left(1 - \frac{\mu_x}{\mu_n} - \frac{\sigma_x^2}{\mu_x \mu_n} + \frac{3\sigma_n^2}{\mu_n^2} - \frac{6\mu_x\sigma_n^2}{\mu_n^3} - \frac{2\sigma_{xn}^2}{\mu_x \mu_n} + \frac{6\sigma_{xn}^2}{\mu_n^2}\right)$$

$$(3.5) \quad E(\hat{\sigma}_x^2) \approx \mu_x - \frac{\mu_x^2}{\mu_n}\left(1 + \frac{\sigma_n^2}{\mu_n^2} + \frac{\sigma_x^2}{\mu_x^2} - \frac{2\,\sigma_{xn}^2}{\mu_x\,\mu_n}\right)$$

Im Fall der obigen Beispieltabelle zeigt sich kein nennenswerter Unterschied zwischen den Binomialvarianzen der Original- und der überlagerten Tabelle. Der Effekt der Überlagerung wirkt sich jedoch bei sehr kleinen Zellenbesetzungen in einer Unterschätzung der Varianz und einer Überschätzung der Anteilswerte aus, wie die Simulationsergebnisse (siehe Tabelle 3.2) belegen.

Tabelle 3.2: Einfluß der Zufallsüberlagerung auf Anteilswert, geschätzte Standardabweichung und geschätzte relative Standardabweichung (Simulationsergebnisse bei tatsächlichem Anteil von 50%)

	Original - Tabelle			Überlagerte Tabelle		
Fall-zahl	Anteil an der Rand-summe in %	Standard-abweichung	relative Std.abw. in %	Anteil an der Rand-summe in %	Standard-abweichung	relative Std.abw. in %
5	50.00	1.5811	31.6228	50.41	1.5372	32.1843
10	50.00	2.2361	22.3607	50.16	2.2232	22.3995
25	50.00	3.5355	14.1421	50.04	3.5328	14.1406
50	50.00	5.0000	10.0000	49.99	4.9986	10.0056
100	50.00	7.0711	7.0711	49.99	7.0707	7.0725
250	50.00	11.1803	4.4721	50.00	11.1800	4.4721
500	50.00	15.8114	3.1623	50.00	15.8113	3.1622

3.2 Statistische Tests

Das Prozentsatzverhältnis (Odds-Ratio) der Beispieltabelle berechnet sich aus: (30*40)/(20*25). Es drückt die relativen Chancenverhältnisse zwischen den beiden Schichten aus, eine weiterführende Schule zu besuchen oder nicht. Da die Überlagerungen der einzelnen Zellen voneinander unabhängig sind, kann für die Berechnung des Erwartungswerts die folgende Näherung verwendet werden (vgl. Anhang):

$$(3.6) \quad E\left(\frac{X_1 X_4}{X_2 X_3}\right) \approx \frac{\mu_{x_1} \mu_{x_4}}{\mu_{x_2} \mu_{x_3}} \left(1 + \frac{\sigma_{x_2^2}}{\mu_{x_2^2}} + \frac{\sigma_{x_3^2}}{\mu_{x_3^2}}\right)$$

Man erhält mit (3.6) für das Odds-Ratio den Erwartungswert von 2,41, der nur geringfügig von dem Wert der Originaltabelle (OR=2,40) abweicht. Die Teststatistik für das Odds-Ratio ist: $t=\ln(OR)/\sigma$, mit: $\sigma=(\Sigma 1/n_{ij})^{1/2}$. Der t-Test für die Originaldaten ergibt t=2,27. In den Simulationen zeigte sich, daß in 5,7 Prozent aller überlagerten Tabellen der t-Test fälschlicherweise keinen signifikanten Zusammenhang ermittelte. Ähnlich reagierte auch der Chi-Quadrat-Unabhängigkeitstest auf die Überlagerung; in 5,2 Prozent der durchgeführten Simulationen wurde der "falsche" Schluß der statistischen Unabhängigkeit gezogen. Diese Ergebnisse hängen natürlich von den kleinen Zellenbesetzungen der Beispieltabelle ab und gelten nur für diese Art der Überlagerung.

Zusammenfassend läßt sich festhalten, daß die Abschätzung der Verzerrung statistischer Kenngrößen um so schwieriger wird, je komplexer diese Größen berechnet werden; dies gilt insbesondere für multivariate Analysen. Zudem sind ohne genaue Kenntnis der statistischen Verteilung des Überlagerungsfehlers in STATIS-BUND kaum exakte Abschätzungen möglich. Diese

Verteilung kann jedoch aus Geheimhaltungsgründen nicht veröffentlicht werden und ist ohnehin mathematisch wesentlich schwieriger zu handhaben als die Normalverteilung. Aus diesen Gründen wenden wir uns im nächsten Abschnitt einem Praxistest zu, bei dem nicht überlagerte und überlagerte Fallzahltabellen aus STATIS-BUND mit den gleichen statistischen Verfahren analysiert werden. Der Vergleich der Ergebnisse kann zeigen, wie sich die Überlagerung mit Zufallsfehlern für die Nutzer von STATIS-BUND auswirkt.

4. Vergleich von multivariaten Analysen mit Original- und überlagerten Fallzahltabellen

Für die Wahl der Testtabellen, mit denen der Vergleich durchgeführt wird, wurden zwei Kriterien zugrunde gelegt. Wie die obige Einführung zeigt, wiegt das Problem der Analyse zufallsüberlagerter Tabellen besonders schwer bei schwach besetzten Zellen. Unter methodischen Gesichtspunkten stellt eine tief gegliederte und spärlich besetzte Tabelle einen 'harten Test' für die Güte der Ergebnisse multivariater Analysen dar. Aus der Nutzerperspektive sollte es sich um eine praxisnahe Anwendung handeln. Einer nach inhaltlichen Kriterien erstellten Tabelle ist gegenüber einer Tabelle mit synthetischen Daten der Vorzug zu geben. Wir haben uns hier an einer Fragestellung aus der schichtspezifischen Bildungsforschung orientiert. Der Schwerpunkt dieser Arbeit ist jedoch methodisch-statistisch. Die inhaltlichen Kriterien für die Wahl der Testtabelle werden im folgenden kurz vorgestellt.

4.1 Bildungschancen in Abhängigkeit von Merkmalen der sozialen Herkunft

4.1.1 Inhaltliche Fragestellung und Beschreibung der Daten

Die Verteilung der Schüler nach sozialer Herkunft auf weiterführende Schulen ist auch nach den Reformversuchen der sechziger Jahre ungleich. Nach den klassischen Schichtindikatoren berufliche Stellung, Bildungsabschluß und Einkommen in der Familie gegliedert, zeigt sich ein Zusammenhang zwischen Schulbesuch/-erfolg und sozialer Schicht: je höher die berufliche Stellung, je höher die Bildungsqualifikation und je höher das Einkommen der Eltern ist, desto größer sind die Chancen, eine weiterführende Schule zu besuchen.

Mit der Variablen berufliche Stellung werden zumeist die Auswirkungen der Arbeits- und Berufserfahrungen der Eltern auf die Erziehungswerte und -handlungen operationalisiert. Es wird angenommen, daß Kinder in Mittelschichtfamilien größere Handlungsspielräume besitzen, die ihre kognitiven Fähigkeiten fördern. Der elterliche Bildungsabschluß ist ein

Indikator für die Vertrautheit der Familie mit den Ausbildungsanforderungen der Schule. Ein hoher Bildungsstatus der Eltern fördert die kindliche Intelligenzentwicklung. Die Einkommenssituation beeinflußt direkt (z.B. Verfügbarkeit von Büchern, Wohnsituation) und indirekt (z.B. das Angewiesensein auf ein frühes Einkommen der Kinder) die Bildungschancen. Seit die Länder keine Daten mehr zu Schulbesuchsquoten nach sozialer Herkunft erheben, ist der Mikrozensus die einzige Quelle der amtlichen Statistik, mit der diese Fragen untersucht werden können. Bildungsforscher kritisieren jedoch die eingeschränkten Analysemöglichkeiten, "... weil die Einteilung der Bevölkerung durch die amtliche Statistik in 4 oder 5 Berufsgruppen sehr undifferenziert und soziologisch nur sehr bedingt brauchbar ist" (Geißler 1987: 88). Daneben fehlt es an aussagekräftigen multivariaten Analysen zu den partiellen Effekten der einzelnen Schichtindikatoren auf die Bildungschancen. Man weiß zwar, daß der Effekt der beruflichen Stellung des Familienvorstandes stärker ist als der Einkommenseffekt (Böttcher 1991: 155, 157), aber gilt dies auch nach Kontrolle des Bildungsniveaus? Wir vermuten, daß kognitive Ressourcen des Elternhauses, vereinfacht gemessen mit dem Bildungsniveau des Familienvorstandes, die Bildungschancen am stärksten beeinflussen.

Wir versuchten die Testtabelle auch nach inhaltlichen Kriterien so zu gestalten, daß ansatzweise die genannten Kritikpunkte überwunden werden können, bzw. Wege aufgezeigt werden, die dem Datenbedarf der Bildungsforschung entgegenkommen. Für die Testtabelle wurden die im Mikrozensus 1987 erfaßten 13 bis 14jährigen Schüler in den Schularten des dreigliedrigen Schulsystems ausgewählt. Diese Altersgruppe eignet sich besonders dafür, die Wirkung der ersten Selektionshürde des deutschen Schulsystems zu untersuchen. [5] Die Testtabelle ist neben dem dichotomen Merkmal Schulart (Hauptschule/weiterführende Schule) nach den Merkmalen Bildungsabschluß des Familienvorstandes, Stellung im Beruf des Familienvorstandes und Einkommen des Familienvorstandes gegliedert. Spalte 1 in Tabelle 4.1 gibt die tatsächliche Anzahl der Schüler je Schulart an. Spalte 2 wurde durch Aggregation der überlagerten Testtabelle gewonnen, während Spalte 3 durch einfache Überlagerung von Spalte 1 entstanden ist.

Aus den Angaben zum allgemeinen und beruflichen Bildungsabschluß des Familienvorstandes wurde eine einfache Bildungsskala erstellt. In Tabelle 4.2 kann man einen fast linearen Zusammenhang zwischen dem Bildungsniveau des Familienvorstands und den Bildungschancen der Kinder erkennen (vgl. Spalte 2). Die insgesamt schwach besetzte Kategorie "Abitur" ist von der Überlagerung stark betroffen; der über die Summe überlagerter Zellen ermittelte Anteilswert (vgl. Spalte 3) liegt um 5,5 Prozentpunkte über dem tatsächlichen Anteilswert. Auf diese Kategorie entfallen bei 42 Zellen

Tabelle 4.1: Anzahl der Schüler nach Schulart, ermittelt als Randsumme der Original-Tabelle, Randsumme der überlagerten Tabelle und überlagerte Randsumme der Original-Tabelle

Schulart	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) Hauptschule	5285	5328	5286
(2) Realschule, Gymnasium	6443	6539	6443
Zusammen	11728	11867	11729

Tabelle 4.2: Schüler in weiterführenden Schulen nach Bildungsabschluß des Familienvorstandes, absolut und als prozentualer Anteil an Schülern aller Schularten

	absolut	in Prozent, ermittelt aus:		
Bildungsabschluß des Familienvorstandes	Randsummen Original-Tabelle	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) Volks-/Hauptschule ohne Lehre, o. Ang.	769	28.1	28.2	28.1
(2) Volks-/Hauptschule mit Lehre	2958	51.7	51.2	51.7
(3) Realschule	1124	77.3	77.9	77.5
(4) Abitur	272	76.0	81.5	74.9
(5) Fachhochschule, Hochschule	1320	90.9	89.8	90.9

Tabelle 4.3: Schüler in weiterführenden Schulen nach beruflicher Stellung des Familienvorstandes, absolut und als prozentualer Anteil an Schülern aller Schularten

	absolut	in Prozent, ermittelt aus:		
Berufliche Stellung des Familienvorstandes	Randsummen Original-Tabelle	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) Un- und angelernte Arbeiter	691	29.0	29.7	29.1
(2) Vor- und Facharbeiter, Meister	1075	46.4	47.3	46.4
(3) Einfache Angestellte und Beamte	1284	68.3	68.5	68.3
(4) Höhere Angestellte und Beamte	1892	85.3	84.7	85.3
(5) Selbständige	985	65.5	64.6	65.4
(6) Nichterwerbstätige, keine Angabe	516	36.1	36.8	35.9

nur 272 Schüler in weiterführenden Schulen. Die große Differenz der Anteile deutet auf Probleme bei der multivariaten Analyse hin. Um den Fehler zu verringern, könnte man rekodieren und neu auszählen. Dies haben wir jedoch nicht getan, da einerseits die Kategorie "Abitur" in der Forschung häufig verwendet wird und andererseits gerade die Folgen der Überlagerung bei geringen Besetzungszahlen untersucht werden sollen.[6]

Zur Differenzierung des Merkmals Stellung im Beruf wurde das Mikrozensusmerkmal Stellung im Betrieb verwendet. Mit den darin enthaltenen Informationen lassen sich die Gruppen der Arbeiter, Angestellten und Beamten entsprechend der Position in der betrieblichen Hierarchie gliedern.[7] Tabelle 4.3 zeigt, daß abgesehen von der heterogenen Gruppe der Selbständigen eine höhere berufliche Stellung mit größeren Anteilen der Kinder in weiterführenden Schulen einhergeht.

In Tabelle 4.4 sind die Anteile der Besucher weiterführender Schulen nach dem Nettoeinkommen des Familienvorstands gegliedert.[8] Ähnlich wie bei der schwach besetzten Kategorie "Abitur" in der Bildungsvariablen kann man für die Einkommenskategorie "unter 1000 DM" eine durch die Überlagerung erzeugte Verzerrung im Vergleich der Anteilswerte feststellen. Die Differenz beträgt hier aber nur drei Prozentpunkte.

Tabelle 4.4: Schüler in weiterführenden Schulen nach monatlichem Nettoeinkommen des Familienvorstandes, absolut und als prozentualer Anteil an Schülern aller Schularten

	absolut	in Prozent, ermittelt aus:		
monatl. Nettoeinkommen des Familienvorstandes	Randsummen Original-Tabelle	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) unter 1000 DM	192	32.4	35.3	32.2
(2) 1000 - 2000 DM	1210	38.8	38.7	38.8
(3) 2000 - 3000 DM	2128	50.9	51.1	51.0
(4) 3000 - 4000 DM	1132	76.4	76.5	76.5
(5) 4000 - 5000 DM	597	86.9	86.3	87.0
(6) 5000 DM und mehr	663	90.6	90.4	90.6
(7) Landwirte, o.Angabe	521	55.4	54.7	55.4

4.1.2 Zum Überlagerungsfehler

In der Testtabelle gibt es K=5*6*7=210 Merkmalskombinationen der 3 unabhängigen Merkmale. n_{k2} sei die Anzahl der Schüler auf weiterführenden Schulen mit Merkmalskombination k und n_k sei die entsprechende Gesamtzahl der Schüler. Die analysierte *Originaltabelle* hat dann die Gestalt:

(4.1) (n_{k2}, n_k), k=1, ... ,K

Die analysierte *überlagerte Tabelle* läßt sich formal als

$$(4.2) \quad (n_{k2} + z_{k2}, n_k + z_k),\ k=1, \ldots, K$$

darstellen, wobei z_{k2} und z_k Realisationen ganzzahliger Zufallszahlen sind. Die analysierte überlagerte Tabelle ist *nicht* identisch mit der von STATIS-BUND gelieferten Tabelle. Die in der Tabelle vorhandenen negativen Zellenbesetzungen müssen aus Plausibilitätsgründen noch nachträglich korrigiert werden. Die für die Logit-Analyse verwendete Tabelle, entstand somit in zwei Schritten:

(i) Bereitstellung der zufallsüberlagerten Tabelle durch STATIS-BUND
(ii) Korrektur unplausibler Fallzahlen durch den Benutzer:
  (a) Negative Fallzahlen werden auf Null gesetzt. [9]
  (b) Die Gesamtzahl der Schüler für eine Merkmalskombination k wird gegebenenfalls soweit erhöht, daß $n_{k2} + z_{k2} < n_k + z_k$ gilt. [10]

Bei einer teilweise so schwach besetzten Tabelle wirkt sich die notwendige Korrektur unplausibler Fallzahlen zusätzlich zu dem STATIS-BUND Überlagerungsfehler aus. Die Kenngrößen für den Gesamtfehler vor und nach der Korrektur sind in Tabelle 4.5 zusammengefaßt.

Tabelle 4.5: Kenngrößen des Gesamtfehlers in der überlagerten Tabelle

Kenngrößen des Gesamtfehlers	vor Korrektur	nach Korrektur	Freiheits-grade
Mittelwert	0.40	0.56	-
Varianz	3.27	3.36	-
Chi-Quadrat-Statistik:			
mit Nullzellen	734	823	419
ohne Nullzellen	185	207	354

Die Chi^2-Statistik wurde zweifach berechnet; einmal mit Berücksichtigung der Nullzellen in der Originaltabelle, die dazu mit dem Wert 0,5 belegt wurden und ein zweitesmal ohne Berücksichtigung der 65 Nullzellen in der Originaltabelle. Beide Ergebnisse weisen auf eine insgesamt gute Anpassung der überlagerten Tabelle an die Originaltabelle hin.

Während die Fallzahlen unserer Testtabelle durch die nachträgliche Korrektur zusätzlich verzerrt werden, ist für die eigentlich interessierenden Anteilswerte eher das Gegenteil festzustellen. Der Anteil an Schülern auf weiterführenden Schulen beträgt in der Originaltabelle 54,94 Prozent, in der überlagerten Tabelle 55,39 Prozent und in der korrigierten Tabelle 55,10 Prozent. Die Werte aus der überlagerten und plausibel gemachten Tabelle

liegen im allgemeinen sehr nahe bei den Originalwerten (siehe Tabellen 4.2-4.4). Sowohl der STATIS-BUND Überlagerungsfehler als auch die Korrektur unplausibler Fallzahlen wirkt sich bei kleinen Fallzahlen stärker aus.

4.2 Vergleichende Analyse

4.2.1 Binäre Logit-Modelle

Bevor wir die Ergebnisse für die Original- und überlagerte Tabelle miteinander vergleichen, stellen wir die Grundzüge der Analyse mit binären Logit-Modellen dar (vgl. Arminger/Küsters 1986). Dabei gehen wir zunächst von der Originaltabelle aus und verwenden folgende Bezeichnungen:

N	:	Stichprobenumfang (hier: N = 11728)
K	:	Anzahl der Merkmalskombinationen der unabhängigen Merkmale (hier: K = 210)
Y	:	Y = 1: Besuch einer Hauptschule Y = 2: Besuch einer weiterführenden Schule
n_k, k=1, ... ,K	:	Anzahl der Fälle mit Merkmalskombination k
n_{k2}, k=1, ... ,K	:	Anzahl der Fälle mit Merkmalskombination k und Y=2
$\pi_k \equiv P(Y=2 \mid K)$	:	Wahrscheinlichkeit für Y=2 bei Vorliegen der k-ten Merkmalskombination

In der Gesamtstichprobe vom Umfang N trete die k-te Merkmalskombination genau n_k-mal auf. Die Zufallsvariable N_{k2}, die die Anzahl der Fälle mit Merkmalskombination k und Y=2 beschreibt, ist dann binomialverteilt:

$$(4.3) \quad P(N_{k2}=n_{k2}) = \binom{n_k}{n_{k2}} \pi_k^{n_{k2}} (1-\pi_k)^{n_k-n_{k2}}$$

Weiterhin wird angenommen, daß der Logarithmus der relativen Chancen $\pi_k/(1-\pi_k)$ eine weiterführende Schule statt eine Hauptschule zu besuchen, linear von den Ausprägungen der unabhängigen Merkmale abhängt:

$$(4.4) \quad \gamma_k = \ln\left(\frac{\pi_k}{(1-\pi_k)}\right) = \beta_0 + \beta_1 X_{k1} + \beta_2 X_{k2} + \ldots + \beta_p X_{kp} = X\beta$$
$$x_{kj} \in \{0,1\}, \ \beta_j \in \mathbb{R}.$$

Das *Nullmodell* M_0 mit nur einem Parameter β_0 ist das einfachste Logit-Modell. Es enthält die Hypothese, daß die Chancen nicht von den erklärenden Merkmalen abhängen. Demgegenüber ist das *saturierte Modell* M_s, das soviele β-Parameter wie Merkmalskombinationen enthält (p+1=K), das komplizierteste, aber am wenigsten informative Modell. Unter der

jeweiligen Modellannahme werden die Maximum-Likelihood-Schätzer (ML-Schätzer) für die ß-Koeffizienten bestimmt. Die zugehörige *Log-Likelihood-Funktion* lautet:

$$(4.5) \quad l(\beta) = \sum_{k=1}^{K} \gamma_k n_{k2} - \log(1+\exp(\gamma_k))\, n_k + \sum_{k=1}^{K} \log\binom{n_k}{n_{k2}}$$

Bei der folgenden Logit-Analyse bleibt der Überlagerungsfehler unberücksichtigt. Die Log-Likelihood-Funktion für die überlagerte Tabelle hat daher die Gestalt (4.5) mit $n_{k2} + z_{k2}$ statt n_{k2} und $n_k + z_k$ statt n_k. Der zugehörige Schätzer für ß wird als *Quasi-Maximum-Likelihood-Schätzer* bezeichnet (vgl. Küchenhoff 1990). In Abschnitt fünf werden Möglichkeiten aufgezeigt, den Zufallsfehler zu berücksichtigen.

Zur Beurteilung der Modellanpassung werden zwei häufig verwendete Kriterien herangezogen, die sich beide aus dem Devianzmaß ableiten. Die *Devianz* ist ein Maß für die Diskrepanz zwischen der unter der Modellannahme geschätzten und der beobachteten Tabelle.

$$(4.6) \quad D(M) = -2\,[\,l(\hat{\beta}_M) - l(\hat{\beta}_{M_S})\,]$$

Dabei bezeichnet $\hat{\beta}_M$ die ML-Schätzung für ß unter dem Modell M. Sollen zwei geschachtelte Modelle M_1 und M_2 miteinander verglichen werden, so bietet sich die Differenz der Devianzen als Teststatistik an, da $D(M_1)-D(M_2)$ unter der Nullhypothese M_1 asymptotisch Chi^2 verteilt ist. $D(M_1)$ bezeichne die Devianz der Individualdaten, das heißt die maximal mögliche Devianz für alle denkbaren Tabellen, und $D(M_0)$ die Devianz des Nullmodells, die maximal mögliche Devianz für die vorliegende Tabelle. Das Bestimmtheitsmaß *Pseudo-R^2*

$$(4.7) \quad R^2(M) = \frac{D(M_0) - D(M)}{D(M_I)}$$

gibt den Anteil an Devianz in den Daten an, der durch die in einem Modell M enthaltenen Effekte erklärt werden kann.

4.2.2 Modellwahl und Modellanpassung

Erhielte man mit überlagerten Fallzahltabellen ein anderes passendes Modell als mit der Originaltabelle, wäre das ein fataler Fehler. Deshalb wird

zunächst geprüft, ob bei einer induktiven Modellsuche gleiche Ergebnisse gefunden werden.

In Tabelle 4.6 sind die unter verschiedenen Modellannahmen ermittelten Kenngrößen sowohl für die Originaltabelle als auch für die überlagerte Tabelle wiedergegeben. Die absoluten Devianzen für die überlagerte Tabelle werden ebenso wie die zugehörigen Freiheitsgrade systematisch überschätzt. Auf den Vergleich von zwei Modellen hat dies jedoch keine Auswirkung, da die Differenz der Devianzen und Freiheitsgrade dicht beieinander liegen und die Chi^2-Tests für beide Tabellen stets zum gleichen Ergebnis, dem Verwerfen der jeweiligen Nullhypothese, führen.

Tabelle 4.6: Kenngrößen verschiedener Logit-Modelle

Modell	Original-Tabelle	Differenz zum Vergleichsmodell	Überlagerte Tabelle	Differenz zum Vergleichsmodell
Modell 0: 1				
Devianz	3107.63		3234.38	
Freiheitsgrade	181		207	
R^2	0		0	
Modell 1: 1+Bild		Modell 0		Modell 0
Devianz	982.06	2125.57	1039.67	2194.71
Freiheitsgrade	177	4	203	4
R^2	0.1317	0.1317	0.1344	0.1344
Modell 2: 1+StiB		Modell 0		Modell 0
Devianz	1037.84	2069.79	1250.69	1983.69
Freiheitsgrade	176	5	202	5
R^2	0.1282	0.1282	0.1215	0.1215
Modell 3: 1+Eink		Modell 0		Modell 0
Devianz	1567.41	1540.22	1700.24	1534.14
Freiheitsgrade	175	6	201	6
R^2	0.0954	0.0954	0.0940	0.0940
Modell 4: 1+Eink+StiB		Modell 7		Modell 7
Devianz	746.81	437.59	905.31	523.56
Freiheitsgrade	170	4	196	4
R^2	0.1462	0.0271	0.1426	0.0321
Modell 5: 1+Eink+Bild		Modell 7		Modell 7
Devianz	643.29	334.07	692.65	310.90
Freiheitsgrade	171	5	197	5
R^2	0.1526	0.0207	0.1557	0.0190
Modell 6: 1+Bild+StiB		Modell 7		Modell 7
Devianz	404.29	95.07	498.83	117.08
Freiheitsgrade	172	6	198	6
R^2	0.1675	0.0058	0.1675	0.0072
Modell 7: 1+Bild+StiB+Eink		Modell 0		Modell 0
Devianz	309.22	2798.41	381.75	2852.63
Freiheitsgrade	166	15	192	15
R^2	0.1733	0.1733	0.1747	0.1747
Maximales R^2	0.1925		0.1981	

Bei großen Fallzahlen sind Chi^2-Tests oft wegen ihrer asymptotischen Eigenschaften wenig informativ für die Beantwortung der Frage, welches Modell die Daten ausreichend gut und sparsam beschreibt. Deshalb war für den Vergleich und die Auswahl eines geeigneten Modells das Bestimmtheitsmaß R^2 eine heuristische Entscheidungsgrundlage. Die daraus abgeleiteten Schlüsse erweisen sich als unabhängig von der Überlagerung. Das maximale R^2 beträgt für die Originaltabelle (überlagerte Tabelle) 19,25 Prozent (19,81 Prozent). Damit enthält die untersuchte Tabelle wesentliche Bestimmungsgründe für die Zielgröße. Das Modell 7 "unabhängiger Einfluß aller drei Einzelmerkmale" mit R^2 = 17,33 Prozent (17,47 Prozent) kann als geeignet akzeptiert werden. Es weicht zwar signifikant von den Daten ab, "erklärt" aber bereits 90,0 Prozent (88,2 Prozent) der maximal erklärbaren Devianz. Diese Entscheidung legt auch der Vergleich weiterer, hier nicht dargestellter Modelle mit Interaktionen der unabhängigen Variablen nahe. Von den drei Einzelmerkmalen hat das Einkommen des Familienvorstandes den geringsten partiellen Einfluß auf die Bildungschance des Kindes. Das partielle R^2 ergibt sich durch einen Vergleich der R^2-Werte für die Modelle 4-6 mit dem Modell 7 (siehe Spalten "Differenz zum Vergleichsmodell" in Tabelle 4.6). Die jeweils in den Modellen 4-6 nicht enthaltene Variable verursacht ein geringeres R^2; der in Spalte "Differenz..." ausgewiesene Wert entspricht dem Nettoeffekt der Variablen. Die überlagerte Tabelle überschätzt im Vergleich zur Originaltabelle die partiellen Effekte Bildungsabschluß (3,21 Prozent vs. 2,71 Prozent) und Einkommen (0,72 Prozent vs. 0,58 Prozent), wohingegen der partielle Einfluß der beruflichen Stellung (1,90 Prozent vs. 2,07 Prozent) unterschätzt wird. Die absoluten Verzerrungen sind jedoch gering und die für die inhaltliche Interpretation wichtige relative Anordnung der drei Merkmale wird durch die Zufallsüberlagerung nicht verändert. Insgesamt kann festgehalten werden, daß die Zufallsüberlagerung die Modellanpassung und die Wahl eines geeigneten Modells nicht wesentlich beeinflußt.

4.2.3 Regressionskoeffizienten

Wir wenden uns nun einem Vergleich der Analyseergebnisse für das ausgewählte Modell 7 zu. Tabelle 4.7 enthält die geschätzten ß-Koeffizienten und ihre geschätzten Standardabweichungen. Richtung und Größenordnung stimmen für überlagerte und Originaltabelle weitgehend überein. Da die Schätzwerte asymptotisch normalverteilt sind, lassen sich näherungsweise 95 Prozent-Konfidenzintervalle für die ß-Koeffizienten der Originaltabelle ableiten. Es zeigt sich, daß nur einer der sechzehn ß-Werte der überlagerten Tabelle außerhalb des Konfidenzintervalls liegt und somit signifikant von dem ß-Koeffizienten der Originaltabelle abweicht. Der Koeffizient zu "Bildungsabschluß = Abitur" wird durch die Überlagerung überschätzt (siehe "Zum Überlagerungsfehler" Abschnitt 4.1.2).

Tabelle 4.7: Vergleich der ß-Koeffizienten und ihrer Standardabweichungen für Modell 7

	Original - Tabelle		Überlagerte Tabelle	
Merkmals-ausprägung	Koeffizient	Standard-abweichung	Koeffizient	Standard-abweichung
1	-1.1842	0.0611	-1.1772	0.0608
Bild(2)	0.6140	0.0559	0.6055	0.0555
Bild(3)	1.3736	0.0832	1.4405	0.0831
Bild(4)	1.3422	0.1387	1.7409	0.1383 *)
Bild(5)	1.8623	0.1180	1.8093	0.1102
StiB(2)	0.4475	0.0664	0.4612	0.0657
StiB(3)	1.0495	0.0737	1.0275	0.0732
StiB(4)	1.3338	0.0911	1.2565	0.0889
StiB(5)	0.8694	0.0852	0.7936	0.0843
StiB(6)	0.1958	0.0803	0.1894	0.0795
Eink(1)	-0.2639	0.1094	-0.2204	0.1057
Eink(2)	-0.1138	0.0533	-0.1321	0.0533
Eink(4)	0.4225	0.0766	0.4495	0.0763
Eink(5)	0.6497	0.1305	0.6732	0.1259
Eink(6)	0.9040	0.1457	0.9994	0.1410
Eink(7)	0.0721	0.0881	0.0476	0.0868

*) ß der überlagerten Tabelle außerhalb des 95% Konfidenzintervalls von ß der Original-Tabelle

Tabelle 4.8: Gegenüberstellung der t-Werte, die zu verschiedenen Ergebnissen bei einem t-Test auf Differenz von zwei ß-Koeffizienten führen

	Original - Tabelle		Überlagerte Tabelle	
Merkmalsausprägung	ß-Differenz	t-Wert	ß-Differenz	t-Wert
Bild(4)-Bild(3)	-0.0314	-0.22	0.3005	2.07
Bild(5)-Bild(4)	0.5201	3.19	0.0684	0.43
StiB(3)-Bild(4)	-0.2928	-1.77	-0.7134	-4.37
StiB(4)-Bild(4)	-0.0084	-0.05	-0.4845	-2.83
StiB(5)-Bild(2)	0.2553	2.31	0.1881	1.72
Eink(4)-Bild(2)	-0.1916	-2.00	-0.1560	-1.64
Eink(6)-Bild(2)	0.2900	1.86	0.3939	2.61
Eink(6)-StiB(4)	-0.4298	-2.36	-0.2571	-1.46

Tabelle 4.8 vergleicht die Ergebnisse bei einem t-Test auf signifikanten Unterschied zwischen je zwei ß-Koeffizienten. Bei einem Signifikanzniveau von fünf Prozent und einem kritischen Wert von ±2 führt die Überlagerung in 8 von 120 Fällen zu einem anderen Ergebnis. Besonders deutlich ist die Diskrepanz in den vier Fällen, in denen der Koeffizient zu "Bildungsabschluß = Abitur" betroffen ist.

Wie schon im Abschnitt 3.1 angesprochen, bewirkt die Fehlerüberlagerung von Zellenbesetzungen tendenziell eine Überschätzung des Anteilswertes von überlagerten Zellen im Vergleich zur Originaltabelle, was sich besonders bei kleinen Fallzahlen deutlich zeigt (siehe Formel 3.3). Dies gilt auch für die durch das Logit-Modell 7 geschätzten Zellenbesetzungen (m_u) der überlagerten Tabelle im Vergleich zu den geschätzten Zellenbesetzungen (m_o) der Originaltabelle.

In Abbildung 4.1 sind die relativen Differenzen geschätzter Zellenbesetzungen $rd_k=(m_{ku}-m_{ko})/m_{ko}$ dargestellt.[11] Eine relative Differenz von Eins zum Beispiel bedeutet, daß bei der überlagerten Tabelle die geschätzte Fallzahl doppelt so groß ist, wie bei der Originaltabelle. Die Verzerrungen werden etwa ab einer geschätzten Fallzahl von zehn zunehmend kleiner und können bei Werten größer 20 praktisch vernachlässigt werden.

Mit Bezug auf Abschnitt 3.1, Formel 3.5, war tendenziell eine Unterschätzung der Varianz zu erwarten. Die in Abbildung 4.2 dargestellte Überschätzung der geschätzten Varianz scheint zunächst damit in Widerspruch zu stehen. Das klärt sich jedoch auf, wenn man beachtet, daß auch die zugrundeliegende geschätzte Fallzahl der überlagerten Tabelle im Vergleich zur Originaltabelle größer ist (siehe Abbildung 4.1).

4.2.4 Residuenanalyse

Weiteren Aufschluß über die Modellanpassung liefert die exploratorische Residuenanalyse, in der die beobachteten Werte mit den geschätzten Werten verglichen werden. Wir verwenden die Pearson-Residuen:

$$(4.8) \qquad S_k = \frac{n_{2k} - \hat{\pi}_k n_k}{\sqrt{\hat{\pi}_k (1 - \hat{\pi}_k)\, n_k}}$$

Ist das Modell gut angepaßt und gilt für jede Merkmalskombination k, daß $n_k \pi_k (1-\pi_k) > 9$ so sind die Pearson-Residuen annähernd standardnormalverteilt. $|S_k| \geq 2$ ist somit ein Indiz für eine durch das Modell schlecht angepaßte Merkmalskombination.

Abbildung 4.1: Relative Differenzen geschätzter Zellenbesetzungen

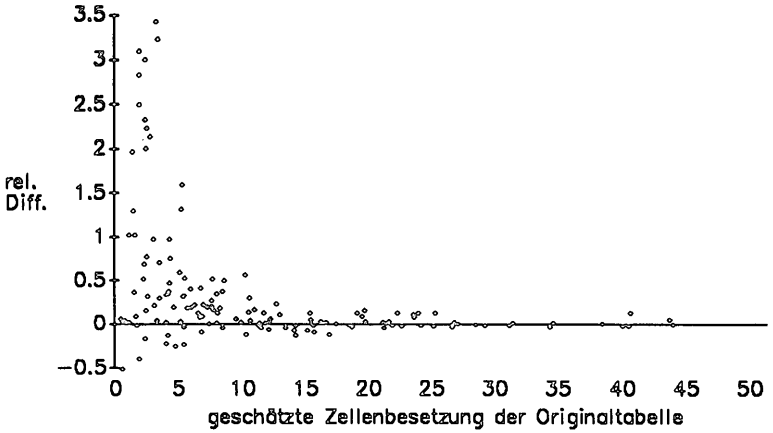
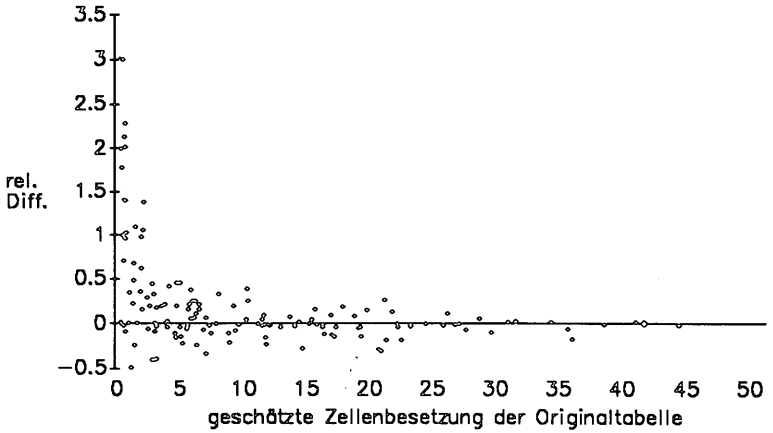


Abbildung 4.2: Relative Differenzen geschätzter Varianzen



In Tabelle 4.9 sind diejenigen Merkmalskombinationen ausgewiesen, die zu unterschiedlichen Ergebnissen bei der Residualanalyse führen. Im Hinblick auf die konservative Bedingung $n_k \pi_k (1-\pi_k) > 9$ wurden in Anlehnung an eine häufig verwendete Faustregel nur Residuen mit $n_k \pi_k \geq 5$ betrachtet. Bei der Originaltabelle lagen elf Residuen im kritischen Bereich; bei der überlagerten Tabelle waren es 16. Ein Vergleich der Residuen in den kritischen Bereichen zwischen Original- und überlagerter Tabelle zeigt jedoch, daß sowohl bei der Originaltabelle Merkmalskombinationen schlecht angepaßt sind, die bei der überlagerten Tabelle gut angepaßt sind, als auch umgekehrt. Insofern kann aus der Residuenanalyse nicht auf einen Überlagerungseffekt geschlossen werden.

Tabelle 4.9: Gegenüberstellung der Pearson-Residuen, die zu verschiedenen Ergebnissen in der Residuenanalyse führen

Merkmalskombination			Original - Tabelle		Überlagerte Tabelle	
Bild	StiB	Eink	geschätzte Fallzahl	Residuum	geschätzte Fallzahl	Residuum
2	2	1	5.66	-0.91	5.43	-2.49
2	3	6	7.20	-0.16	8.11	-2.51
2	4	6	26.92	1.49	26.99	2.44
2	5	6	44.61	-1.75	44.02	-2.22
2	5	7	156.81	-2.23	148.55	-1.44
3	4	2	6.43	-3.05	4.80	1.23
3	5	2	16.56	1.13	14.32	2.32
3	6	7	9.80	1.13	9.96	2.08
4	5	1	2.05	1.18	5.30	-2.03
4	6	2	9.51	1.21	11.06	2.51
5	3	6	7.46	-2.07	7.48	-0.68

4.2.5 Direkter Test auf Unterschiede zwischen den Tabellen

Um den Einfluß der Überlagerung auf die geschätzten Effektkoeffizienten im Modell 7 direkt zu testen, wurde eine zusammengesetzte Tabelle gebildet. Diese neue Tabelle besitzt als viertes Merkmal die Tabellenart mit zwei Ausprägungen (TAB = 1: Originaltabelle, TAB = 2: überlagerte Tabelle).

Tabelle 4.10: Direkter Test auf Überlagerungseffekt

Modell	Devianz	Freiheits-grad	R^2 in %
1) 1+Tab+Bild+Stib+Eink	697.83	373	17.38
2) 1+Tab+Bild+Stib+Eink +Tab*(Bild+Stib+Eink)	690.97	358	17.40

Modell 1 enthält die Hypothese, daß der (geschätzte) Einfluß der drei unabhängigen Merkmale auf die Chance eines Schülers, eine weiterführende Schule zu besuchen, nicht von der Tabellenart abhängt. Im Gegensatz dazu wird im zweiten Modell unterstellt, daß dieser Einfluß sehr wohl von der Tabellenart und damit von der Überlagerung abhängt. Es zeigt sich, daß die 15 zusätzlichen Koeffizienten eine nicht signifikante Devianzreduktion von 6,86 erzielen. Ein t-Test für die einzelnen Koeffizienten ergibt ferner, daß lediglich der Koeffizient zu "Bild = Abitur" * "Tab = 2" signifikant von Null abweicht. Die geringfügige Verringerung der Devianz im zweiten Modell ist also im wesentlichen auf eine Verbesserung der Schätzungen für diese kritische Kategorie zurückzuführen.

5. Möglichkeiten der Einbeziehung eines Zufallsfehlers

Im letzten Kapitel wurde eine Logit-Analyse einer STATIS-BUND Tabelle durchgeführt, die beim Schätzalgorithmus die Zufallsüberlagerung völlig außer acht ließ (Quasi-ML-Schätzung). In diesem Kapitel werden drei Möglichkeiten aufgezeigt, wie der Überlagerungsfehler berücksichtigt werden könnte. Die Vorschläge behandeln *exakt normalverteilte Fehler*. Da für eine konkrete Tabelle aus STATIS-BUND die Normalverteilungsannahme verletzt ist, lassen sich daraus keine allgemeingültigen Empfehlungen für die praktische Arbeit ableiten.

5.1 Fehlerverringerung mittels linearer Regression

Da die Summen fehlerüberlagerter Zellen stärker um die wahren Randsummen streuen als die ausgezählten und überlagerten Randsummen, liegt es nahe, eine Tabelle zu schätzen, die in sich möglichst konsistent ist.[12)] Diese geschätzte Tabelle ließe sich in einem zweiten Schritt für die eigentlichen Analysen verwenden. Zur Fehlerverringerung könnte die folgende einfache Methode dienen, mit der berücksichtigt wird, daß zwischen den Zellen und den Randverteilungen eine lineare Beziehung besteht. Anhand der Beispieltabelle 3.1 aus Kapitel 3 soll diese Methode beschrieben werden. Die abhängige Variable Y besteht aus den vier überlagerten Zellhäufigkeiten, je zwei überlagerten Spalten- bzw. Zeilensummen sowie der überlagerten Gesamtsumme. Die vier zu schätzenden Originalfallzahlen bilden die Regressionskoeffizienten $\beta = (\beta_1, \ldots, \beta_4)$. Unter der Annahme, daß Y normalverteilt ist mit $E(Y) = X*\beta$, wobei die (9x4) - Designmatrix X nur aus Nullen oder Einsen besteht, kann eine lineare Regression durchgeführt werden. Bei einem exakt normalverteilten Fehler konnte der Gesamtfehler in den vier überlagerten Fallzahlen verringert werden (vgl. Tabelle A1 im Anhang).

Für STATIS-BUND Tabellen ist der Erfolg dieses Verfahrens gering, da die Normalverteilungsannahme nicht erfüllt ist. Dennoch wurde das Verfahren auf die in Kapitel 4 beschriebene Testtabelle angewendet. Da bei der Testtabelle sehr viele schwach besetzte Zellen vorliegen, wird man mit dem Problem geschätzter negativer Zellenbesetzungen konfrontiert sein. Um dies zu umgehen, wurde die überlagerte Testtabelle nachträglich über das Einkommensmerkmal aggregiert. Dieses Vorgehen bietet zugleich einen empirischen Test, wie sich eine Aggregation überlagerter Tabellenfelder auswirkt. Zum Vergleich wurde eine weitere Tabelle ausgezählt, in der, wie in der nachträglich aggregierten Testtabelle, das Einkommensmerkmal unberücksichtigt blieb, jedoch sind hier die Zellen nur einfach mit Zufallsvariablen überlagert. Die nachträglich aggregierte Testtabelle (Schulbesuch * Bildungsabschluß * berufliche Stellung) weicht mit einem Chi-Quadrat Wert von 393 bei 59 Freiheitsgraden signifikant von der aggregierten Originaltabelle ab. Folglich weichen auch die Analysen dieser Tabelle gravierend von den Ergebnissen der Originaltabelle ab. Wendet man das beschriebene Verfahren an, reduziert sich der Chi-Quadrat Wert auf 43, d.h. die geschätzte Tabelle weicht nicht signifikant von der Originaltabelle ab. In einem Fall wurde eine negative Zellenbesetzung geschätzt. Die vor der Überlagerung über das Einkommensmerkmal aggregierte Tabelle weicht von der Originaltabelle nur um einen nicht signifikanten Chi-Quadrat Wert von 8 ab. Nach Anwendung des Verfahrens auf diese Tabelle zeigt sich jedoch der eingeschränkte Nutzen des Verfahrens, denn die geschätzte Tabelle ist mit einen Chi-Quadrat Wert von 11 schlechter als die überlagerte Tabelle.

5.2 Logit-Modell mit Fehlern in den Fallzahlen

Im folgenden wird ein binomiales Logit-Modell für Tabellen mit normalverteilten Fehlern formuliert und dem binomialen Logit-Modell aus Abschnitt 4.2.1 gegenübergestellt. Es wird angenommen, daß die Randsumme W_k normalverteilt ist ($W_k \sim N(n_k, \sigma^2)$) und die Fallzahl W_{k2} die unabhängige Summe aus einer binomialverteilten Größe ($B(n_k, \pi_k)$) und einer normalverteilten Größe ($N(0,\sigma^2)$) ist. Die Aspekte der Ganzzahligkeit und Positivität von Fallzahlen werden in diesem Modell außer acht gelassen:

(a) Logit-Modell mit normalverteilten Überlagerungsfehlern unbekannte Parameter: $\beta = (\beta_0, \ldots \beta_p)^t, \pi_k, n_k, \sigma^2$

(5.1) $W_{12}, \ldots, W_{K2}, W_1, \ldots, W_K$ *stochastisch unabhängig mit*

(i) $P\{W_{k2} \le t\} = \sum_{i=0}^{n_k} N_{(i,\sigma^2)}(-\infty, t]\, B_{(n_k,\pi_k)}(i), \quad t \in \mathbb{R}$

(ii) $P\{W_k \le t\} = N_{(n_k,\sigma^2)}(-\infty, t), \quad t \in \mathbb{R}$

(iii) $\pi_k = \exp(X_k^t \beta) \;/\; (1 + \exp(X_k^t \beta))$

(b) Logit-Modell ohne Überlagerungsfehler unbekannte Parameter:
$\beta = (\beta_0, \ldots \beta_p)^t$, π_k

(5.2) $N_{12}, \ldots, N_{K2}$ *stochastisch unabhängig mit*

(i) $P\{N_{k2} \leq t\} = B_{(n_k, \pi_k)}(-\infty, t]$, $t \in \mathbb{R}$

(ii) $\pi_k = \exp(X_k^t \beta) \; / \; (1 + \exp(X_k^t \beta))$

Unter der einfachsten Modellannahme $\pi = \pi_k = \exp(\beta_0)/(1+\exp(\beta_0))$ vergleichen wir die beiden Likelihood-Gleichungen, die sich aus der Log-likelihood-Funktion durch Differentiation nach β_0 ergeben. Gleichungen für die Parameter n_k und σ^2 im Fall (a) sind nicht aufgeführt. Phi(i, σ^2) bezeichnet die Dichte der Normalverteilung mit Mittelwert i und Varianz σ^2.

(5.3)

$$a) \quad \sum_{k=1}^{K} \sum_{i=0}^{n_k} \frac{f_{ki}(\pi)}{f_k(\pi)} [i - n_k \pi] = 0 \quad mit$$

$$f_k(\pi) = \sum_{i=0}^{n_k} f_{ki}(\pi) = \sum_{i=0}^{n_k} \varphi_{(i, \sigma^2)}(w_{k2}) \binom{n_k}{i} \pi^i (1-\pi)^{n_k - i}$$

$$b) \quad \sum_{k=1}^{K} [n_{k2} - n_k \pi] = 0$$

Hieraus ist bereits ersichtlich, welche rechentechnischen Schwierigkeiten beim Lösen der Likelihood-Gleichungen einer zufallsüberlagerten Tabelle mit normalverteilten Fehlern im Vergleich zum Standardmodell auftreten können. Lösungsansätze für Maximum-Likelihood-Schätzungen von gemischten Verteilungen, wie sie die Verteilung von W_{k2} darstellt, finden sich zum Beispiel in Everitt/Hand (1981).

Die in der Literatur diskutierten 'Fehler-in-den-Variablen-Modelle' behandeln Regressionsprobleme, bei denen die erklärenden Variablen fehlerbehaftet sind (vgl. Bickel/Ritov 1987, Küchenhoff 1990). Küchenhoff (1990) befaßt sich unter anderem mit Logit-Modellen der Form $P(Y=1 \mid X=x)=(1+\exp(-\alpha-\beta x))^{-1}$, wobei die normalverteilte Einflußgröße X nicht beobachtbar ist, da sie von einer ebenfalls normalverteilten Störgröße überlagert wird. Ein verwandtes Problem ist unter dem Stichwort "overdispersion" bekannt (vgl. McCullagh/Nelder 1983). Es tritt insbesondere dann auf, wenn wichtige erklärende Merkmale nicht verfügbar sind oder eine zu hohe Aggregationsstufe gewählt wurde. Die Erfolgswahrscheinlichkeit π_k in Kategorie k wird dann als zufällig angesehen. Pierce/Sands (1975) untersuchten zum Beispiel ein Logit-Modell der Form $\pi_k = P(Y=1 \mid x_k) = (1+\exp(x_k^t \beta + \sigma z))^{-1}$, wobei hier die Störgröße z nicht beobachtbar ist. Ist z normalverteilt, so besitzt π_k eine

logistisch-normale Verteilung (vgl. Williams 1982). Aus der Literatur sind uns Logit-Modelle mit Fehlern gemäß (5.1) nicht bekannt.

5.3 Modifizierte Maximum-Likelihood-Schätzung

Offensichtlich ist es nicht einfach, ein statistisches Verfahren zu finden, das den Zusatzfehler in multivariaten Analysen überlagerter Fallzahltabellen geeignet berücksichtigt. Im folgenden werden erste, pragmatisch orientierte Überlegungen in diese Richtung vorgestellt.

In Abschnitt 3.1 wurde gezeigt, daß der Anteilswert bei überlagerten Fallzahlen tendenziell um den Faktor σ_n^2/μ_n^2 überschätzt wird (siehe Formel 3.3). Um diese Verzerrung zu korrigieren, verringern wir die interessierenden Fallzahlen n_{k2} um diesen Faktor. Mithilfe einer weiteren Näherung (Taylor-Reihe) läßt sich zeigen, daß die Korrektur unter der Normalverteilungsannahme Verbesserungen bringt, wenn die Randsumme größer gleich fünf ist.[13] Man erhält mit (5.4) die neuen Fallzahlen n^*_{k2}:

$$(5.4) \quad n^*_{k2} = n_{k2} \left(1 - \frac{\sigma_n^2}{n_k^2}\right)$$

Im Gegensatz zur Überschätzung des Anteilswertes wurde im Abschnitt 3.1 eine Unterschätzung der Varianzen nach dem Binomialansatz festgestellt (siehe Formeln 3.4 und 3.5). Da die ML-Schätzung für Logit-Modelle mittels einer iterativen gewichteten linearen Regression erfolgt, bei der die geschätzten Varianzen als Gewichtungsfaktoren eingehen, ist eine Korrektur dieser Varianzen naheliegend. Mit m_{k2} seien die geschätzten Fallzahlen bezeichnet; die Gesamtschülerzahl sei n_k. Die Varianzfunktion für normale Logit-Modelle lautet, der Einfachheit halber ohne Indizes: $var(m) = m - m^2/n$. In Anlehnung an Formel 3.5 korrigieren wir mit (5.5) die iterativ berechnete Varianz durch Vertauschen der Vorzeichen in der Klammer:[14]

$$(5.5) \quad \hat{\sigma}^2_{m_{k2}} = m_{k2} - \frac{m_{k2}^2}{n_k} \left(1 - \frac{\sigma^2_{n_k}}{n_k^2} - \frac{\sigma^2_{m_{k2}}}{m_{k2}^2}\right)$$

Beobachtungen mit hoher Varianz werden bei der Parameterschätzung weniger berücksichtigt als Beobachtungen mit geringer Varianz. Die modifizierte Varianzfunktion (5.5) ist plausibel, da die Varianz besonders bei kleinen Zellenbesetzungen künstlich größer wird. Die Varianz der Überlagerung wird mit Drei angenommen. Der Einfachheit halber wird diese Annahme auch für die Varianz der geschätzten Fallzahlen m getroffen. Eventuell durch die Modellierung entstehende Kovarianzen zwischen den

geschätzten Zellenbesetzungen und der Randsumme bleiben dabei ebenfalls außer Acht.

Nach diesen beiden Korrekturen zeigen die Ergebnisse in Tabelle 5.1 eine deutliche Verbesserung im Vergleich zu den vorherigen Ergebnissen (siehe Tabellen 4.6 und 4.7).

Tabelle 5.1: Modifizierte ML-Schätzung für Modell 7

Devianz: 342.04
Freiheitsgrade: 192

Merkmals-ausprägung	Koeffizient	Standard-abweichung
1	-1.1931	0.0614
Bild(2)	0.6050	0.0562
Bild(3)	1.3991	0.0845
Bild(4)	1.5384	0.1476
Bild(5)	1.7595	0.1153
StiB(2)	0.4709	0.0666
StiB(3)	1.0632	0.0742
StiB(4)	1.3326	0.0914
StiB(5)	0.8424	0.0861
StiB(6)	0.2154	0.0812
Eink(1)	0.4293	0.0773
Eink(2)	0.5609	0.1298
Eink(4)	0.8763	0.1466
Eink(5)	-0.2970	0.1113
Eink(6)	-0.1280	0.0537
Eink(7)	0.0139	0.0887

Gegenüber einer Devianz für Modell 7 bei der überlagerten Tabelle (Originaltabelle) von 381,75 (309,22) erhält man nach der Modifikation eine Devianz von 342,04. Die ß-Werte sind entsprechend näher an den Ergebnissen der Originaltabelle. Insbesondere weicht nun der Parameter für Bildungsabschluß = "Abitur" nicht mehr signifikant vom Schätzwert der Originaltabelle ab (vgl. Tabellen 4.7 und 5.1). Bei den t-Tests der Differenzen von je zwei ß-Werten sind nur noch drei Differenzen kritisch; zuvor waren es acht Differenzen. Ähnliche Ergebnisse erhält man beim Residuenvergleich.

Berücksichtigt man, daß die Modifikationen nur eine erste, grobe Annäherung an die schwierige Frage der Behandlung von Fehlern in den beobachteten Fallzahlen bei multivariaten Analysen darstellen, sind die Verbesserungen überraschend gut. Dies könnte ein Anlaß sein, in dem Bereich der Modifikation von Anteilswerten und Binomialvarianzen weiter nach Lösungen zu suchen.

6. Schluß

STATIS-BUND bietet neben einer Vielzahl von Zeitreihen und Strukturdaten unter bestimmten Voraussetzungen die Möglichkeit, per Online-Anschluß schnell und flexibel Auswertungen aus Einzeldaten der amtlichen Statistik zu erhalten. Dieser Zugriff erfolgt in Form von frei spezifizierbaren Fallzahltabellen, die aus Anonymisierungsgründen mit Zufallszahlen überlagert werden. Mit dem vorliegenden Aufsatz wurden erste Untersuchungen über die Auswirkungen dieser Zufallsüberlagerung auf multivariate Analysen vorgestellt. Zusammenfassend läßt sich festhalten, daß selbst bei einer sehr spärlich besetzten Testtabelle keine gravierenden Verzerrungen multivariater Analyseergebnisse auftreten. Die Wahl eines statistischen Modells, das die Struktur der Tabelle hinreichend gut beschreibt, wird durch die Überlagerung nicht wesentlich beeinflußt. Das gleiche gilt für die relative Bedeutung der Einzelmerkmale, wenngleich ihr partieller Einfluß auf die Zielgröße unter- bzw. überschätzt wird. Wie zu erwarten, sind Einzelergebnisse, die sich aus sehr schwach besetzten Kategorien ableiten, stark verzerrt. In unserer Beispieltabelle ist dies die Kategorie "Abitur" des Merkmals Bildungsabschluß des Familienvorstands. Als Präventivmaßnahme schlagen wir vor, die Randsummen der Einzelmerkmale stets auszuzählen. Weicht die überlagerte Randsumme stark von der Summe der überlagerten Einzelfelder ab, so ist eine erneute Tabellenerstellung unter Zusammenfassung kritischer Kategorien zu empfehlen; auch wenn dies unter inhaltlichen Gesichtspunkten nicht immer leicht fallen mag.

Bei der (bewußt) so tief gegliederten Testtabelle ist wegen der Größe des Stichprobenfehlers die statistisch gesicherte Interpretation der Analyseergebnisse teilweise nicht mehr gegeben. Weitere Logit-Analysen mit stärker besetzten STATIS-BUND-Tabellen zeigten keine nennenswerten Verzerrungen.[15] Dennoch lassen sich diese Befunde nicht ohne weiteres verallgemeinern, da hierzu weitere Analysen mit anderen Verfahren (lineare Regression, Log-lineare Modelle etc.) und anderen Tabellen erforderlich sind.

Neben der Darstellung eines praxisorientierten Vergleichs von Logit-Analysen wurden in Abschnitt 3.1 Näherungsformeln entwickelt, die Hinweise auf Art und Größenordnung der Verzerrungen wichtiger Schätzwerte liefern. Darüber hinaus wurden in den Abschnitten 5.2-5.3 erste Ansätze vorgestellt, wie man Fehler in den Fallzahlen bei multivariaten Analysen angemessen berücksichtigen könnte. In diesem Bereich ist weitere Forschung nötig.

Anmerkungen

1) Wir danken Siegfried Gabler für hilfreiche Hinweise zu mathematischen Fragen.
2) In Zukunft wird auch ein direkter Datentransfer von einem *Liefer-PC* zu dem PC eines Interessenten über das Telefonnetz möglich sein. Der Kunde sendet dazu über seinen mit einem Telefonmodem ausgestatteten PC die Datenanforderung an den PC des Statistischen Bundesamtes, der dann seinerseits automatisch die Bereitstellung der gewünschten Daten veranlaßt.
3) Aus Geheimhaltungsgründen kann die genaue Verteilung der Zufallszahlen nicht veröffentlicht werden.
4) Dabei wurde von einem durchschnittlichen Zuschlagsfaktor von k=1,6 des Standardfehlers für den Design-Effekt ausgegangen. Mit diesem Faktor wird berücksichtigt, daß sich der Standardfehler durch die Klumpenauswahl des Mikrozensus gegenüber einer reinen Zufallsstichprobe im allgemeinen erhöht.
5) Besucher von Sonderschulen werden im Mikrozensus wie Hauptschüler erfaßt. Gesamtschüler blieben außer Acht. Die Daten des Mikrozensus 1987 beziehen sich auf die Bevölkerung am Familienwohnsitz.
6) Bei den folgenden Analysen berücksichtigen wir die Vergrößerung des Stichprobenfehlers durch den Design-Effekt nicht (siehe Anmerkung 4). Es muß außerdem darauf hingewiesen werden, daß die statistische Aussagefähigkeit dieser Testtabelle aufgrund der geringen Besetzungszahlen zum Teil gering ist. Streng genommen können die Analysen nur heuristischen Charakter haben.
7) Auf eine Trennung zwischen Angestellten und Beamten wurde verzichtet.
8) Da für Landwirte im Mikrozensus keine Einkommensangabe erhoben wird, aber eine vollständige Gliederung angestrebt wurde, sind die Landwirte mit der Gruppe ohne Angabe bzw. ohne Einkommen zusammengefaßt.
9) Eine nachträgliche Korrektur negativer Fallzahlen ist nicht nötig, wenn man in STATIS-BUND die Option "Fallzahlen = positiv" wählt. Die Option bewirkt quasi, daß die Korrektur bereits intern bei der Erstellung der Tabelle durchgeführt wird. In jedem Falle ist aber der gesamte Überlagerungsfehler für kleine Fallzahlen tendenziell positiv.
10) Diese Korrektur ist nicht erforderlich, wenn man (z.B. in GLIM; Payne 1985) eine logistische Regression schätzt. Statt der Vektoren Besucher weiterführender Schulen (n_{k2}) und Gesamtschülerzahl ($n_{\cdot}$) im Logit-Modell wird in der logistischen Regression ein Fallzahlvektor Schulbesuch (Hauptschüler | Besucher weiterführender Schulen) je Merkmalskombination spezifiziert. Da jedoch die Varianz des Zufallsfehlers für die daraus zu berechnenden Gesamtschülerzahlen vergrößert wird, wurde das Logit-Modell gewählt.
11) Berücksichtigt wurden nur Zellen, die sowohl in der überlagerten als auch in der Originaltabelle mit Werten größer als Null besetzt waren. Der Übersichtlichkeit halber sind in den Abbildungen nur relative Differenzen für geschätzte Fallzahlen bis 50 ausgewählt worden.
12) Es handelt sich dabei um ein Problem der Anpassung von Zellen an vorgegebene Randverteilungen. Das Verfahren des Iterative Proportional Fitting (IPF) kann dafür nicht verwendet werden, da es von Fehlern ausgeht, die proportional zur Zellenbesetzung sind. Diese Voraussetzung wird jedoch von der Überlagerung im STATIS-BUND nicht erfüllt.
13) Tatsächlich wurden jedoch für die Testtabelle auch bei schwächer besetzten Randsummen Verbesserungen festgestellt.
14) Entstehen bei der Modifikation der Zellenbesetzungen negative Werte, werden sie auf Null gesetzt. Im Schätzalgorithmus muß aus diesem Grund ein Wert größer Null vergeben werden; hier 0,5.

15) Einer der Vergleiche wurde mit der über das Einkommensmerkmal aggregierten Tabelle durchgeführt (siehe Abschnitt 5.1). Bei einem anderen Test verwendeten wir eine Tabelle zur Fragestellung des Erwerbsstatus von Frauen in Abhängigkeit von der Kinderzahl. Diese Tabelle wies mit einer Dimension von 3*5 Zellen in der am schwächsten besetzten Zelle eine Fallzahl von 91 auf.

Literatur

Arminger, G./Küsters, U., 1986: Statistische Verfahren zur Analyse qualitativer Variablen (=Forschungsbericht der Bundesanstalt für Straßenwesen, Bereich Unfallforschung, 147). Bergisch-Gladbach.

Bickel, B.J./Ritov, Y., 1987: Efficient Estimation in the Errors in Variables Model. The Annals of Statistics, 15, 2:513-540.

Böttcher, W., 1991: Soziale Auslese im Bildungswesen. Die Deutsche Schule, 83, 2: 151-161.

Christensen, R., 1990, Log-linear Models, New York: Springer.

Everitt, B.S./Hand, D.J., 1981: Finite Mixture Distributions. London: Chapman and Hall.

Geißler, R., 1987: Soziale Schichtung und Bildungschancen, S. 79-110, in: ders. (Hrsg.): Soziale Schichtung und Lebenschancen in der Bundesrepublik Deutschland, Stuttgart: Enke.

Küchenhoff, H., 1990: Logit- und Probitregression mit Fehlern in den Variablen (=Mathematical systems in Economics, Vol. 177). Frankfurt: Hain.

Kühn, J./Pfrommer, F./Schrey, E., 1984: Zur technischen Weiterentwicklung des Statistischen Informationssystems. Wirtschaft und Statistik, 12:981-987.

McCullagh, P./Nelder, J.A., 1983: Generalized Linear Models. London: Chapman and Hall.

Payne, C.D., 1985: The GLIM Systems. Release 3.77. Oxford: NAG.

Pierce, D.A./Sands, B.R., 1975: Extra-Bernoulli variation in binary data. Technical Report 46, Dept. of Statistics, Oregon State University.

Statistisches Bundesamt, 1992: Statistisches Informationssystem des Bundes. Datenbestand 1992/1993. Wiesbaden.

Wauschkuhn, U., 1982: Anpassung von Stichproben und n-dimensionalen Tabellen an Randbedingungen. München/Wien: Oldenbourg.

Williams, D.A., 1982: Extra-Binomial Variation in Logistic Linear Models. Applied Statistics, 31: 144-148.

Anhang

1. Bestimmung der Erwartungswerte einer Funktion durch Taylor-Reihen

X und Y seien Zufallsvariablen mit Mittelwert $\mu(x)$ bzw. $\mu(y)$ und Varianz $\sigma^2(x)$ bzw. $\sigma^2(y)$. $\mu(x)$ sei dabei die Häufigkeit in der interessierenden Zelle der Originaltabelle und $\mu(y)$ die entsprechende Randsumme. Der Erwartungswert einer Funktion f(X,Y) läßt sich näherungsweise mit Hilfe einer Taylor-Reihen-Entwicklung (unter entsprechenden Voraussetzungen an f) bis zur zweiten Ordnung berechnen. Es gilt:

$$(A.1) \quad E(f(X,Y)) \approx f(\mu_x,\mu_y) + \frac{1}{2}\sigma_x^2 f_{xx}(\mu_x,\mu_y) + \frac{1}{2}\sigma_y^2 f_{yy}(\mu_x,\mu_y) + \sigma_{xy}^2 f_{xy}(\mu_x,\mu_y)$$

Hieraus lassen sich unmittelbar die Formeln (3.3)-(3.5) ableiten. Die Näherungsformel (3.6) für das Odds-Ratio ergibt sich mit $X=X_1X_4$, $Y=X_2X_3$ und $f(X,Y)=X/Y$, wenn die überlagerten Werte X_1, X_2, X_3, X_4 als unabhängig angenommen werden. In diesem Fall gilt nämlich:

(A.2) $\mu_{(x_1 \cdot x_2)} = \mu_{x_1} \cdot \mu_{x_2}$

(A.3) $\sigma^2_{(x_1 \cdot x_2)} = \mu^2_{x_1}\sigma^2_{x_2} + \mu^2_{x_2}\sigma^2_{x_1} + \sigma^2_{x_1}\sigma^2_{x_2}$

2. Fehlerverringerung mittels linearer Regression

Wir verwenden für die Darstellung der Methode eine Realisation der Überlagerung der Beispieltabelle 3.1 mit Zufallsvariablen. Tabelle A1 zeigt die abhängige Variable, den Aufbau der Designmatrix und die Schätzergebnisse. Wie man im Vergleich der Chi^2-Werte sieht, kann dadurch insgesamt eine gute Anpassung der überlagerten Tabellenfelder an die Originalwerte erzielt werden. Im Einzelfall (siehe Zeile 1) kann die Schätzung aber auch schlechtere Ergebnisse bringen.

Tabelle A1: Verringerung des Gesamtfehlers mittels linearer Regression (Tabelle 3.1 als Original-Tabelle)

Tabellen-feld	Schätzung	Fallzahlen Original	Überlagert	Design-Matrix X			
n_{11}	30.28	30.00	30.22	1	0	0	0
n_{12}	24.97	25.00	25.49	0	1	0	0
n_{21}	19.85	20.00	18.93	0	0	1	0
n_{22}	39.48	40.00	39.14	0	0	0	1
$n_{1.}$	55.25	55.00	54.43	1	1	0	0
$n_{2.}$	59.33	60.00	59.37	0	0	1	1
$n_{.1}$	50.13	50.00	51.61	1	0	1	0
$n_{.2}$	64.45	65.00	65.35	0	1	0	1
$n_{..}$	114.58	115.00	113.99	1	1	1	1
Chi-Quadrat							
Zeilen 1-4:	0.0105	-	0.0870				
Zeilen 1-9:	0.0256	-	0.1621				